



大模型推荐技术及展望

冯福利 博士

fengfl@ustc.edu.cn

2023/8

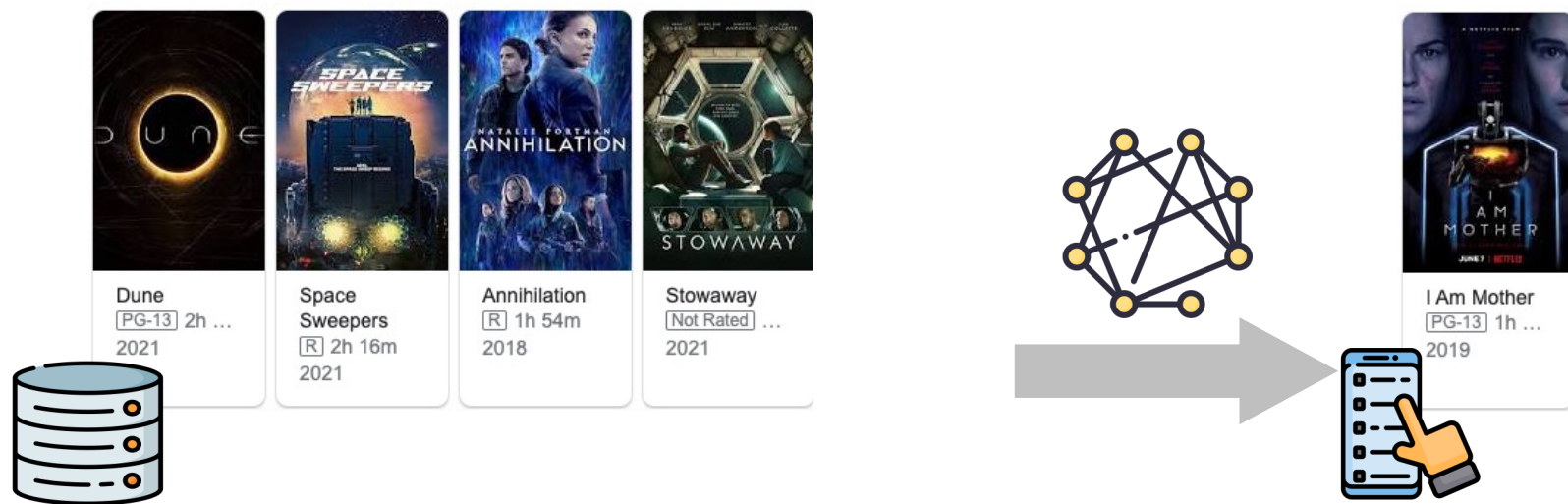
- 1. 推荐及LLM简介**
2. LLM赋能推荐系统
3. 大模型推荐展望

推荐方法的本质

□ 本质：拟合历史用户行为数据，预测未来用户行为

阶段1：在历史数据里学

阶段2：预测用户下一个喜欢的物品



- User：行为多样、模式复杂，受众多外界因素影响
- Item：item间众多低频关联，不断出现新item

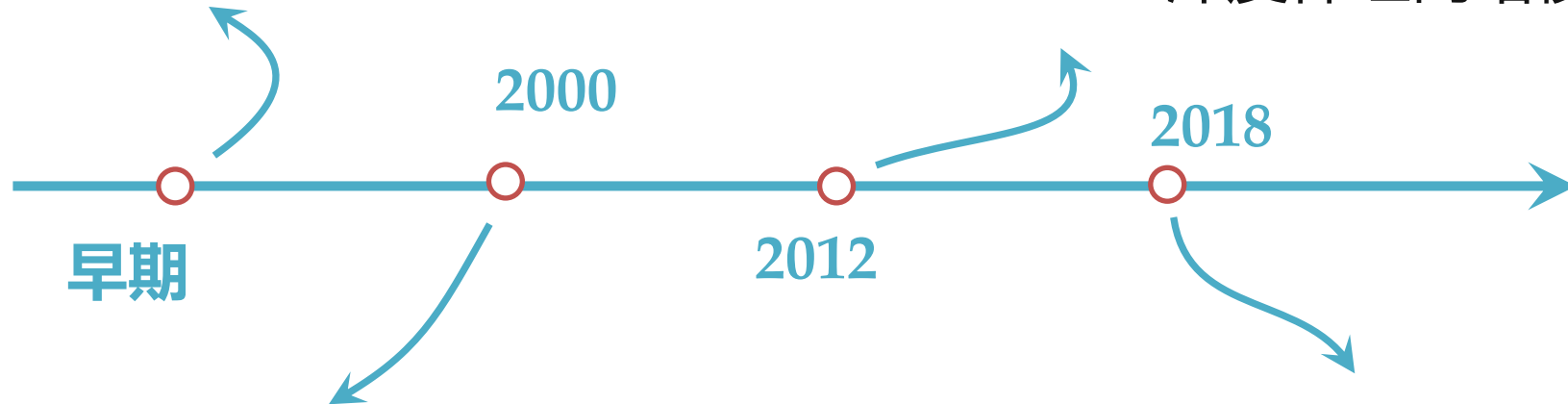
理解不到位，泛化能力差，推荐不满意

大模型新范式

自然语言处理（NLP）领域的发展：

NLP 1.0: 字典/词汇表 + 规则

NLP 3.0: 深度神经网络模型



NLP 2.0: 统计模型

NLP 4.0: 预训练 + 微调 + Prompt 模式
进入大型语言模型 (LLM) 时代

智能涌现



用简明的语句概括一下新闻的大概内容：
“新华社电...”



答案是：这篇新闻大概讲述了“阿根廷夺得世界杯冠军的历程”

学会了！

泛化能力

学会举一反三



帮我概括一下
NCF文章的主要内容

好的，NCF利用神经网络对协同过滤信息建模...



英文中的“Good Morning”是中文的早上好

学会了！“Good Morning”翻译成早上好

规划能力

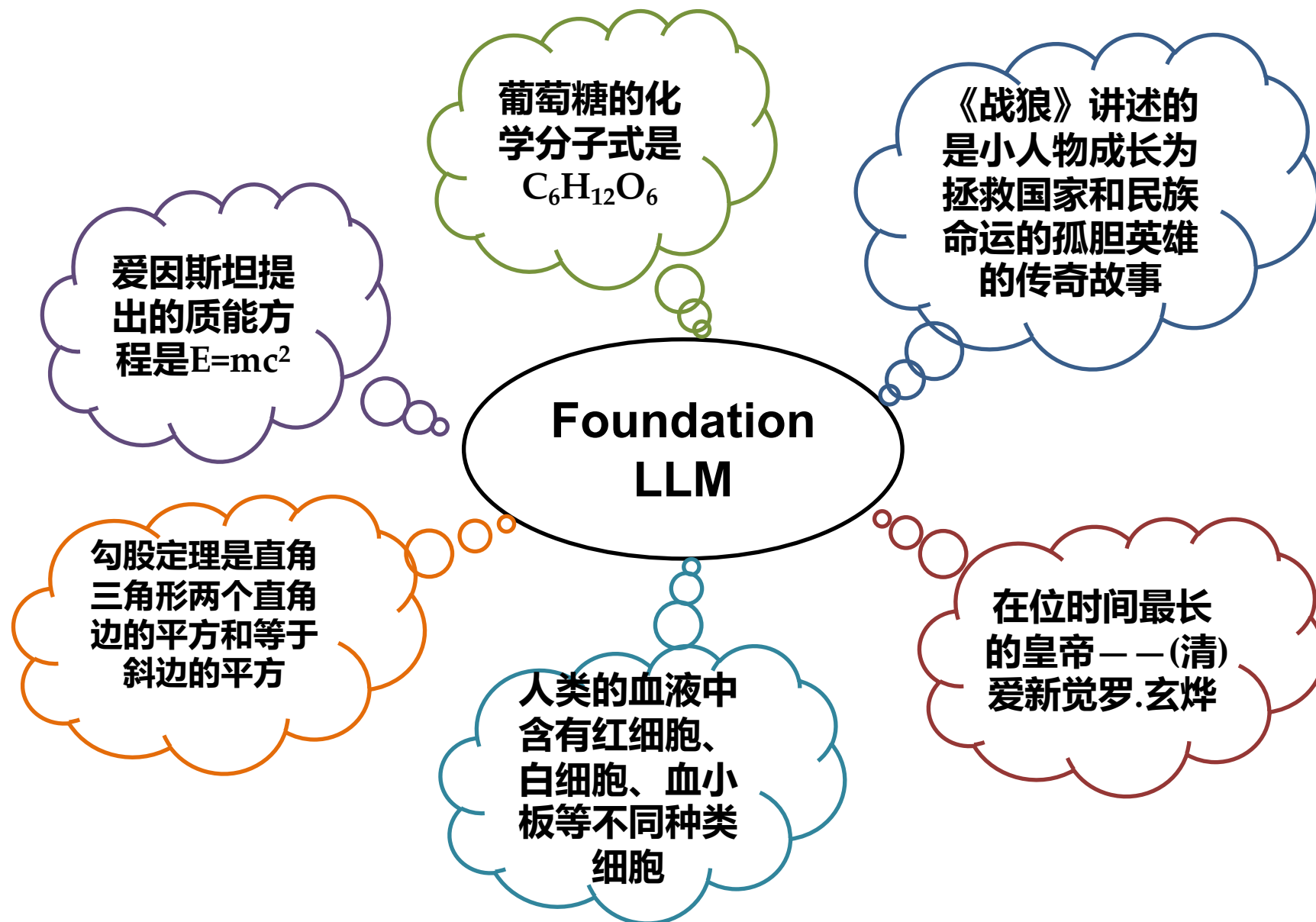
切分简化任务



帮我用英文概括一下这个中文博客
“大语言模型...”

规划：先用中文概括，再翻译成英文—“This blog is about the LLM...”

知识模型



推荐系统 “馋” 大模型啥？

- User：行为多样、模式复杂，受众多外界因素影响
 - Item：在物理世界中存在千丝万缕的关联，很多关联都很低频
- 理解不到位，泛化能力差，推荐不满意

• **Learning**：Pretrain-finetune, prompt learning, instruction-tun.

• **Model**：Well-trained models with extraordinary abilities

Representation：

Textual feature, text is all you need

Chat

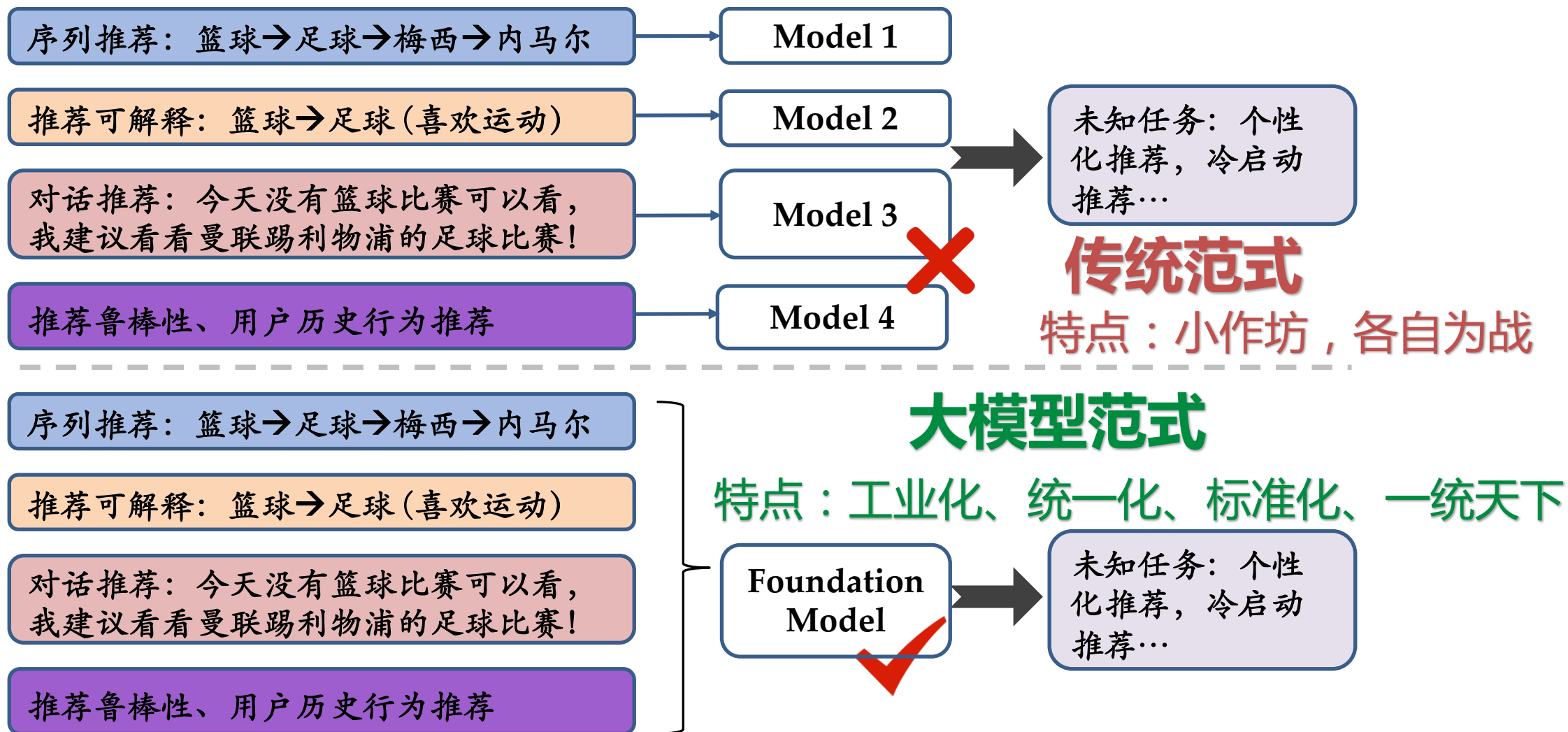
Generalization & Generation：

Few/zero-shot, cross-domain, knowledge, personalized gen.

• **Architecture**：Transformer, self-attention

推荐系统“馋”大模型啥？

□ Open-ended Domains and Tasks



推荐系统 “馋” 大模型啥？

Open-ended domains & tasks

□ 用文本统一表示，用prompt统一task，用LLM跨域

- **Learning** : Prompt learning/instruction-tuning [1,2,3,5,6,9,12]
- **Representation** : Text is all you need [4,5,6,7,8,9,12]
- **Generalization** : Few-shot, cross-domain [5,6,7,8], know. [10,11]

[1] Zhang Yuhui, et al. "Language Models as Recommender Systems: Evaluations and Limitations" *NeurIPS Workshop* 2021.

[2] Zhang Zihuo and Wang Bang. "Prompt Learning for News Recommendation" *SIGIR* 2023.

[3] Geng Shijie et al. "Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)" *RecSys* 2022.

[4] Li Jiacheng et al. "Text Is All You Need: Learning Language Representations for Sequential Recommendation" *KDD* 2023.

[5] Cui Zeyu et al. "M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems" *arXiv* 2022.

[6] Zhang Junjie et al. "Recommendation as Instruction Following: A Large Language Model Empowered Recommendation Approach" *arXiv* 2023.

[7] Gao, et al. "Chat-REC: LLMs-Augmented Recommender System" 2023.

[8] Liu, et al. "Is ChatGPT a good recommender? A preliminary Study" 2023.

[9] Bao Keqin et al. "LLM4Rec: Large Language Models for Recommendation via A Lightweight Tuning Framework" *RecSys* 2023.

[10] Xi Yunjia et al. "Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models" *arXiv* 2023.

[11] Li Jinming et al. "GPT4Rec: A Generative Framework for Personalized Recommendation and User Interests Interpretation" *arXiv* 2023.

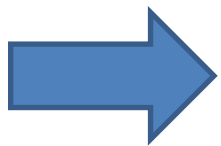
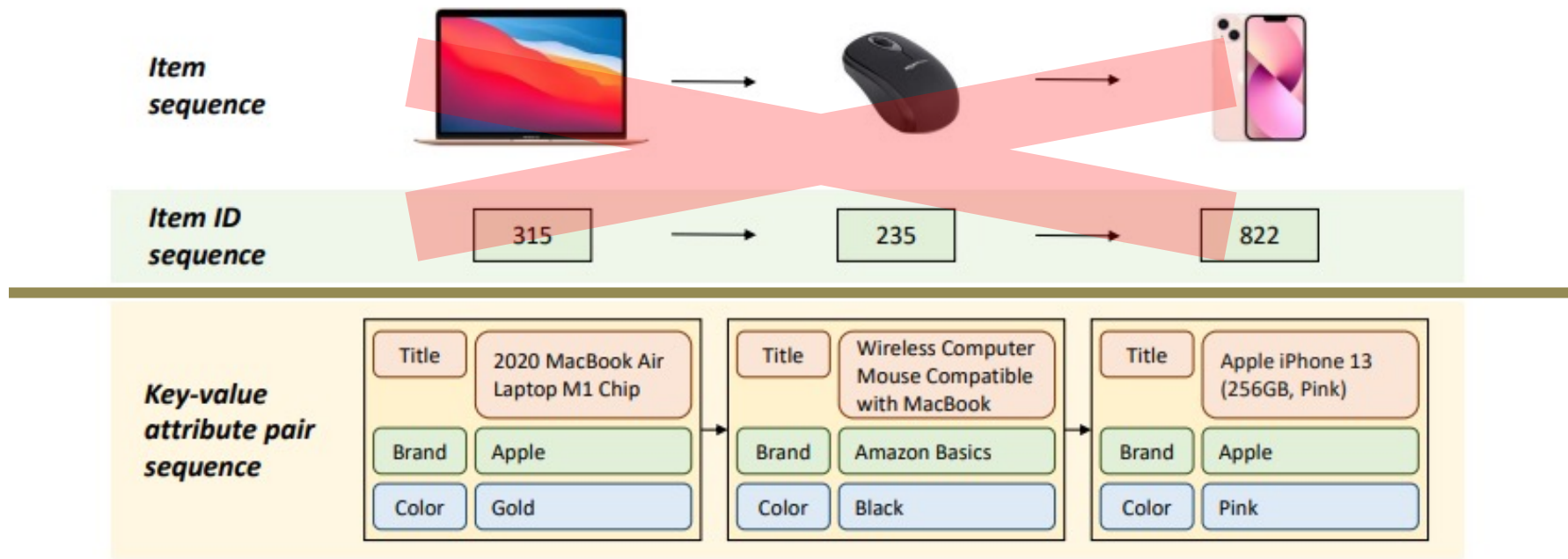
[12] Hou Yupeng et al. "Large Language Models are Zero-Shot Rankers for Recommender Systems" *arXiv* 2023.

[13] Bao Keqin et al. "A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems" *arXiv* 2023.

1. 推荐及LLM简介
2. LLM赋能推荐系统
3. 大模型推荐展望

LLM赋能推荐---Representation

□ Text is all you need (NO item ID)



Use nature language to do recommendation

-> Low resource, cold start ...

LLM赋能推荐---Representation

□ Text is all you need (NO item ID)

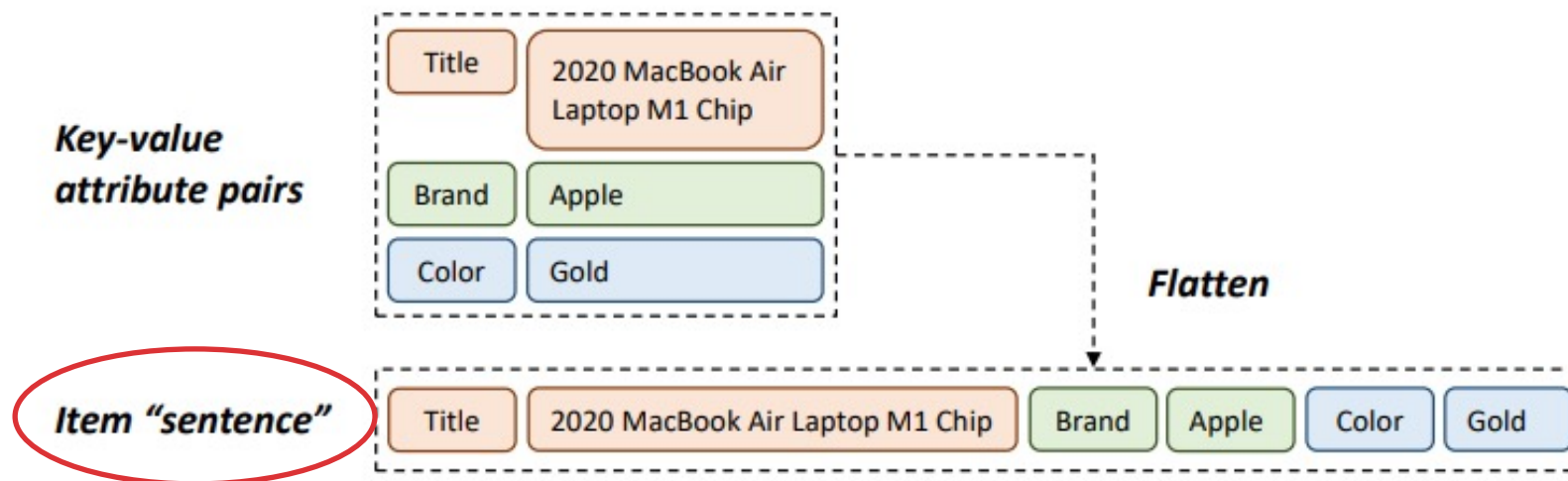


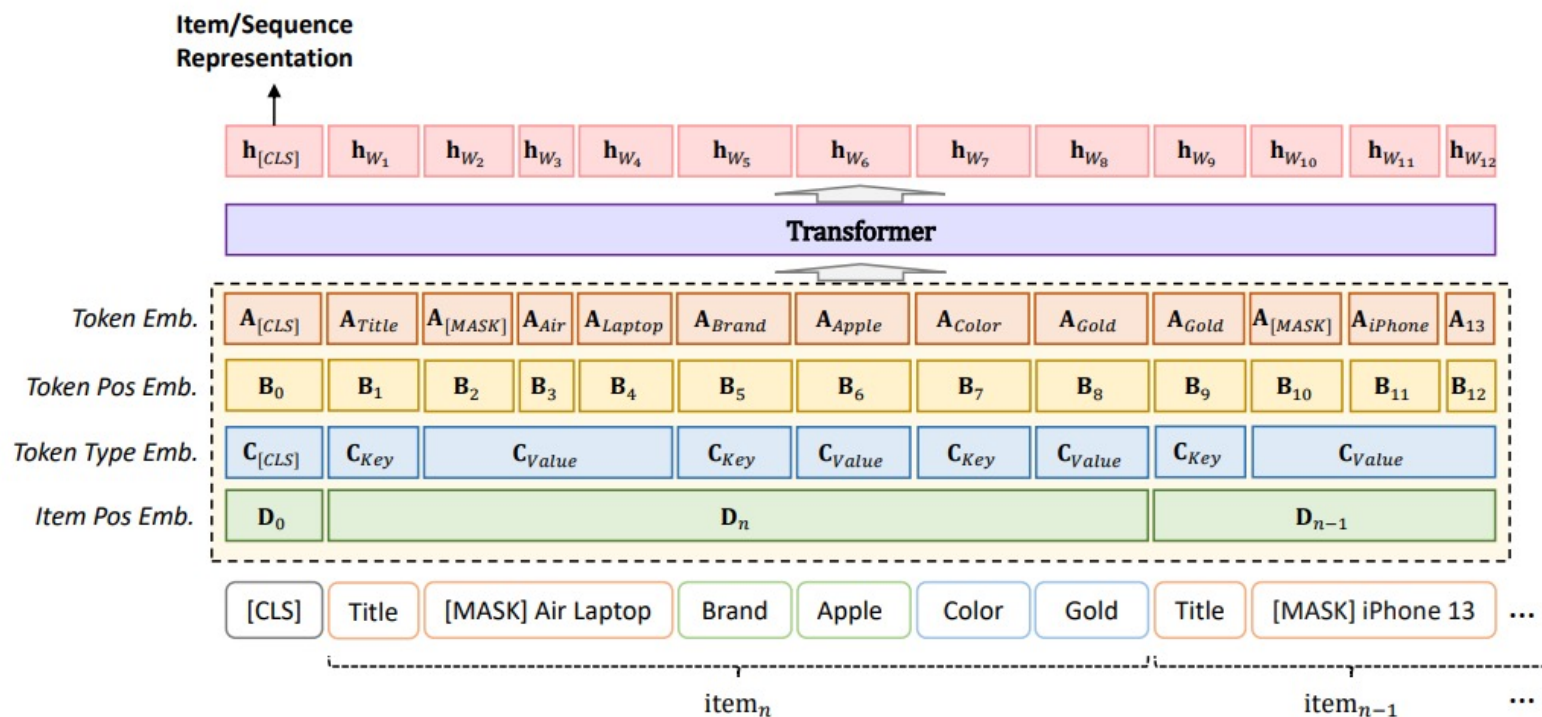
Figure 2: Model input construction. Flatten key-value attribute pairs into an item "sentence".

Item → Sentence

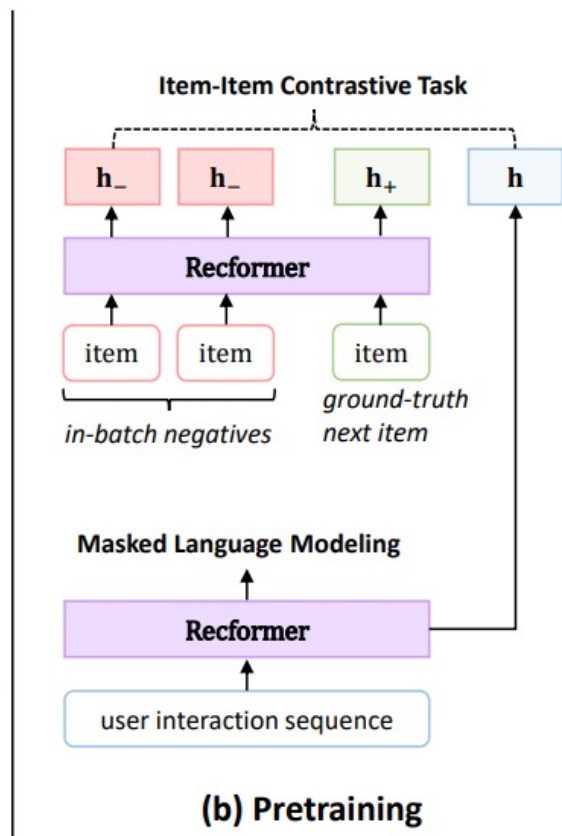
Item sequence → Long sentence

LLM赋能推荐---Representation

□ Text is all you need (NO item ID)



(a) Recformer Model Structure



(b) Pretraining

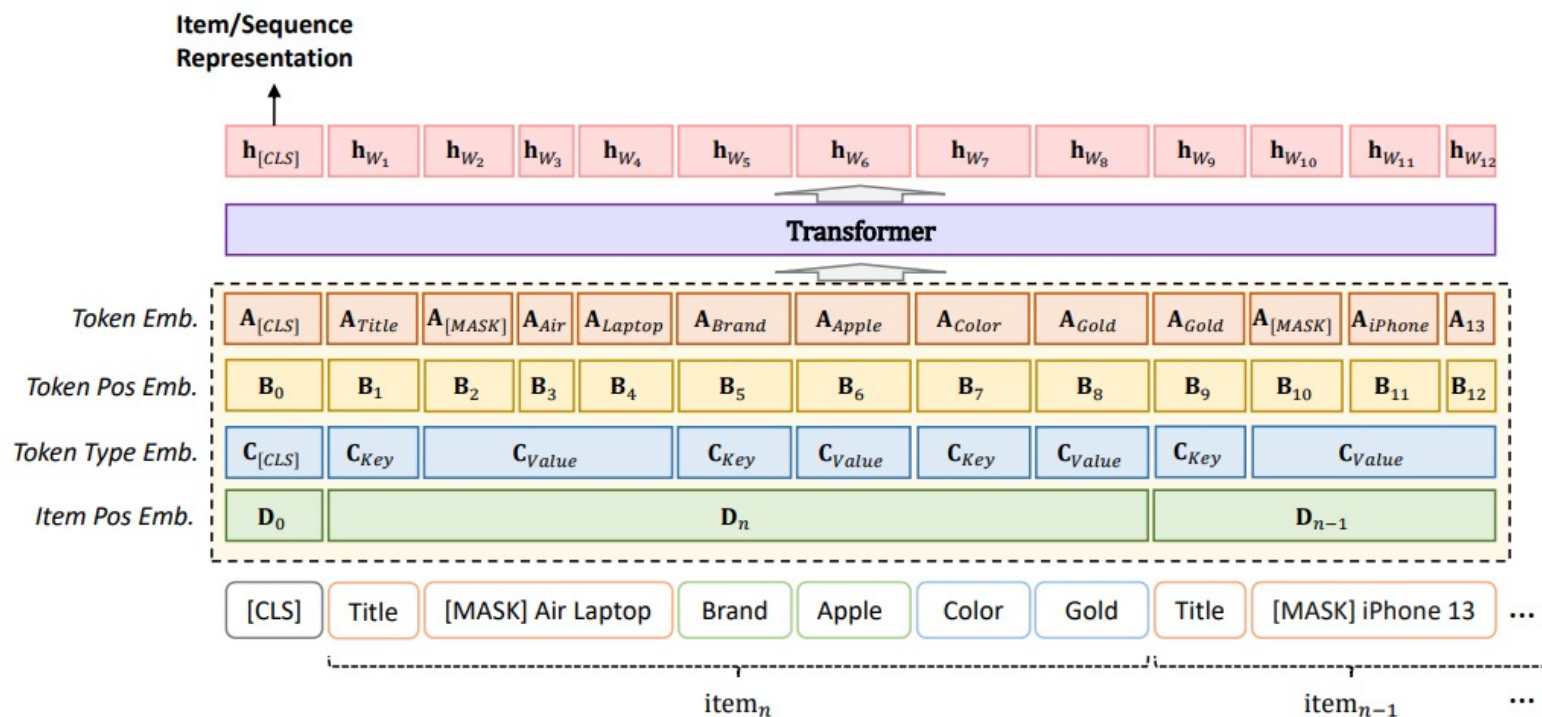
$$E_w = \text{LayerNorm}(A_w + B_w + C_w + D_w)$$

$$[h_{[CLS]}, h_{w_1}, \dots, h_{w_l}] = \text{Longformer}([E_{[CLS]}, E_{w_1}, \dots, E_{w_l}])$$

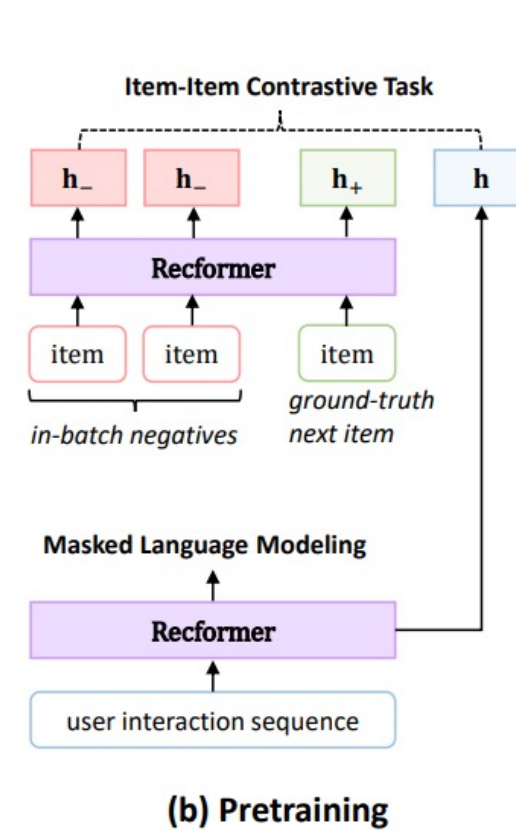
Li Jiacheng et al. "Text Is All You Need: Learning Language Representations for Sequential Recommendation" KDD 2023.

LLM赋能推荐---Representation

□ Text is all you need (NO item ID)



(a) Recformer Model Structure



(b) Pretraining

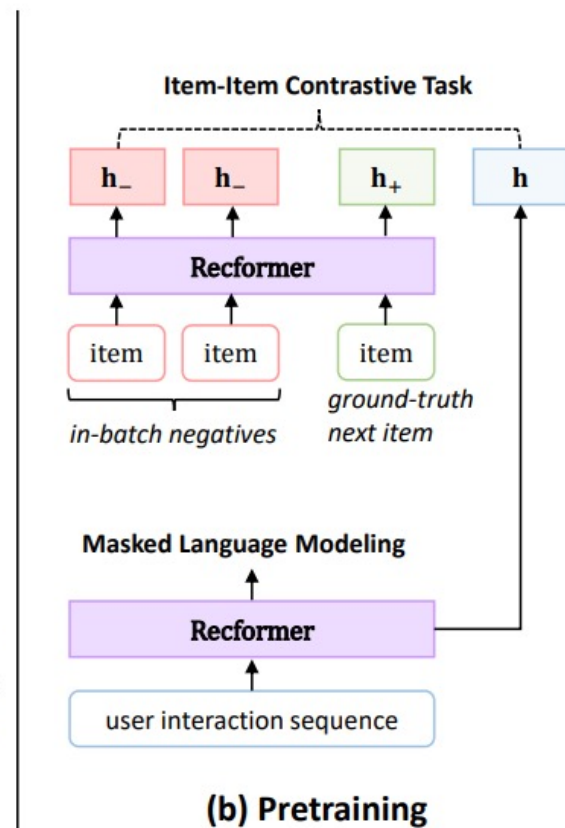
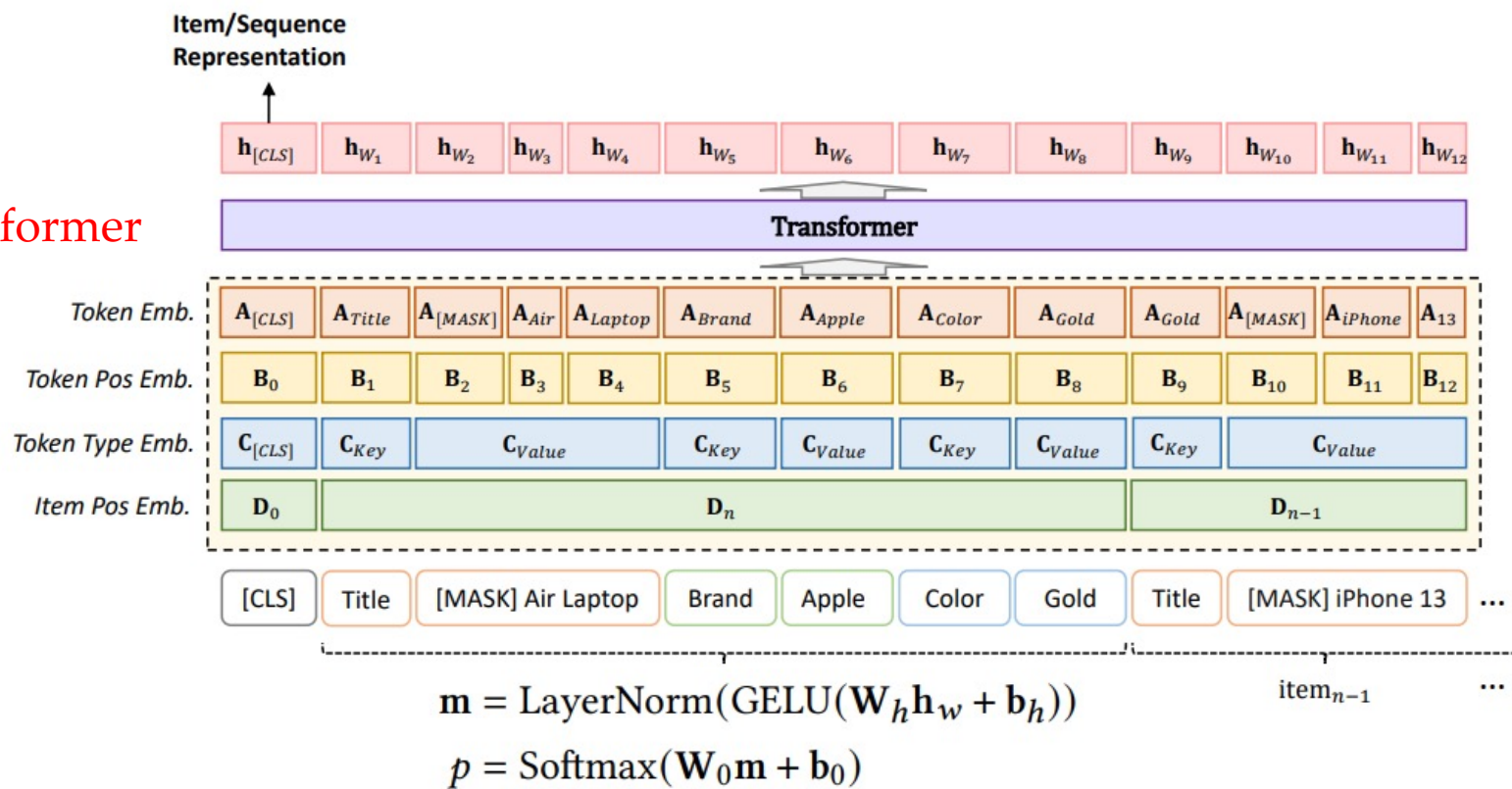
Prediction:
$$r_{i,s} = \frac{\mathbf{h}_i^\top \mathbf{h}_s}{\|\mathbf{h}_i\| \cdot \|\mathbf{h}_s\|}$$

Li Jiacheng et al. "Text Is All You Need: Learning Language Representations for Sequential Recommendation" KDD 2023.

LLM赋能推荐---Representation

□ Text is all you need (NO item ID)

Longformer



Pre-train:

$$\mathcal{L}_{\text{MLM}} = - \sum_{i=0}^{|\mathcal{V}|} y_i \log(p_i)$$

$$\mathcal{L}_{\text{IIC}} = - \log \frac{e^{\text{sim}(\mathbf{h}_s, \mathbf{h}_i^+) / \tau}}{\sum_{i \in \mathcal{B}} e^{\text{sim}(\mathbf{h}_s, \mathbf{h}_i) / \tau}}$$

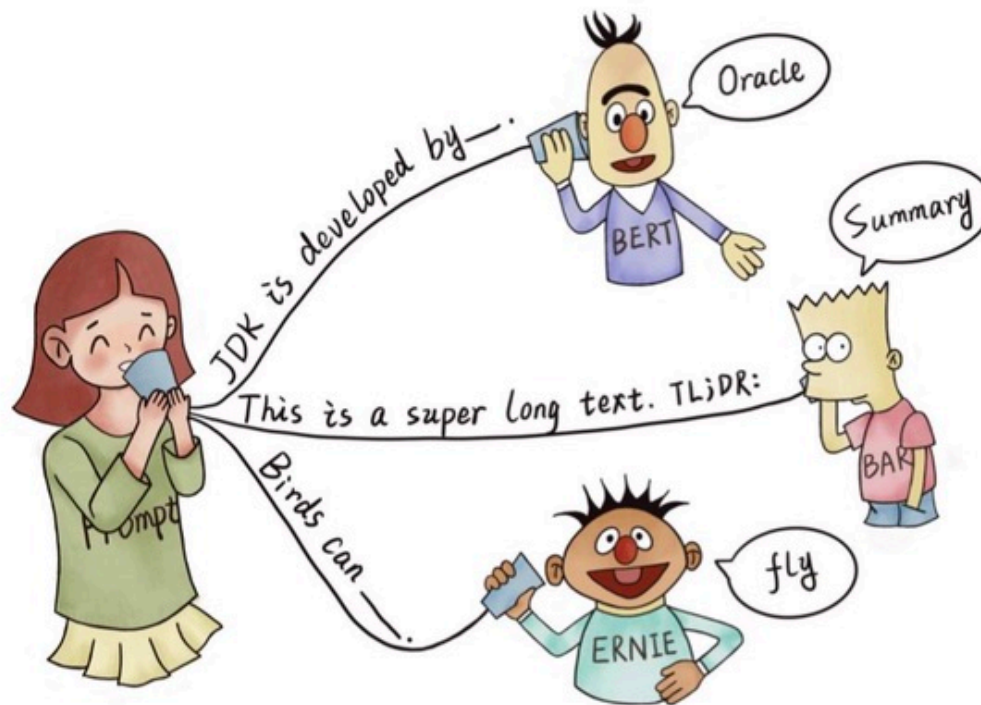
$$\mathcal{L}_{\text{PT}} = \mathcal{L}_{\text{IIC}} + \lambda \cdot \mathcal{L}_{\text{MLM}}$$

Li Jiacheng et al. "Text Is All You Need: Learning Language Representations for Sequential Recommendation" KDD 2023.

LLM赋能推荐---Learning Paradigm

□ Prompt learning

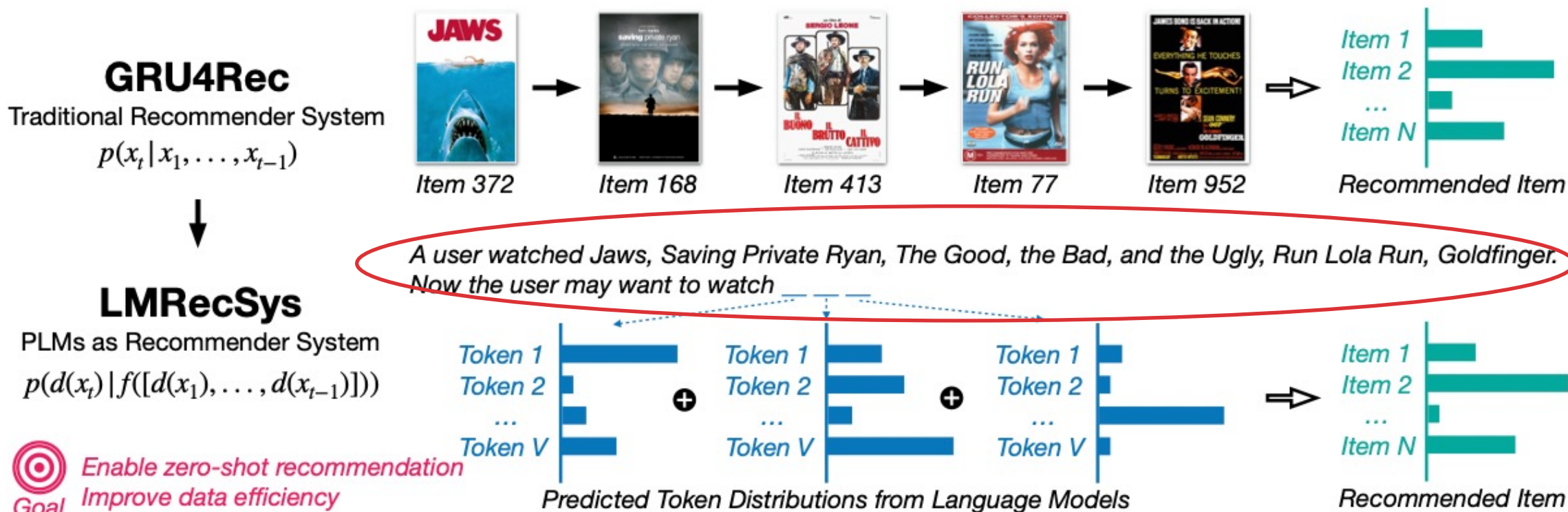
每个task被一个prompt描述
Sentiment analysis :
Input : I love this movie.
Prompt: This is [z].



Sample-level learning $f(y|x)$ $\rightarrow$ Task-level learning $f(y|x,p)$

LLM赋能推荐---Learning Paradigm

□ Prompt learning



BERT/GPT2

LLM赋能推荐---Generalization

□ Open-ended domains and tasks

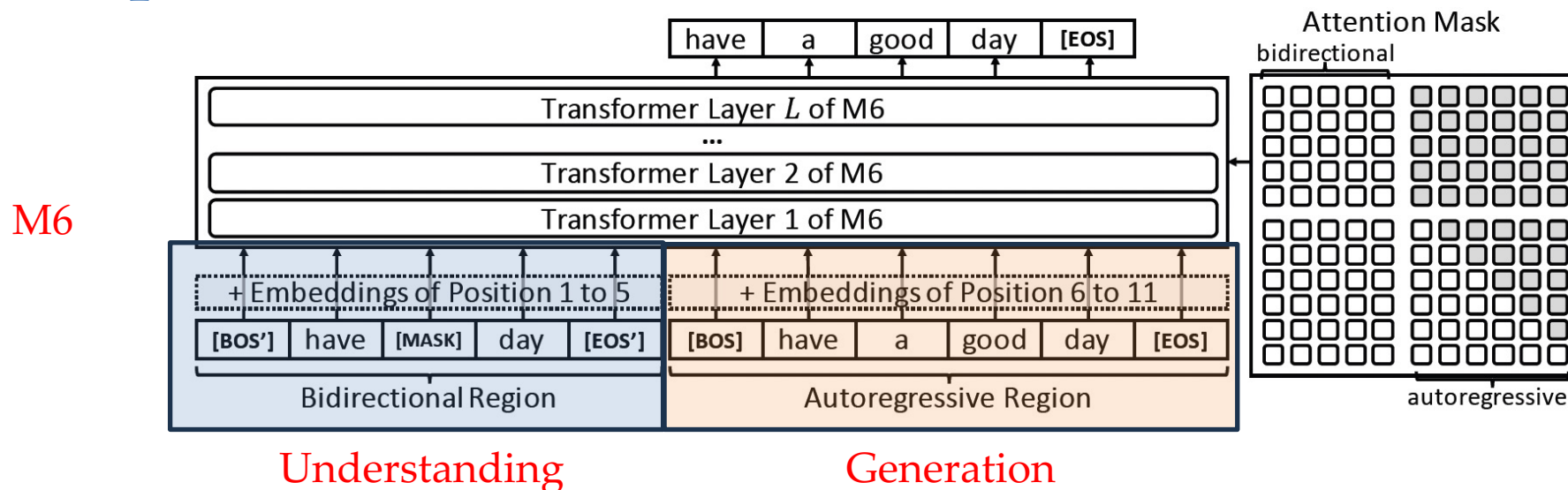


Figure 1: The text infilling objective used by the pretraining procedure of M6. [MASK] represents an undetermined number of unknown tokens. [BOS] and [EOS] mean the beginning and the end of a sentence, respectively. The autoregressive language modeling loss is imposed on the outputs of the autoregressive region, and not on the bidirectional re-

LLM赋能推荐---Generalization

□ Open-ended domains and tasks

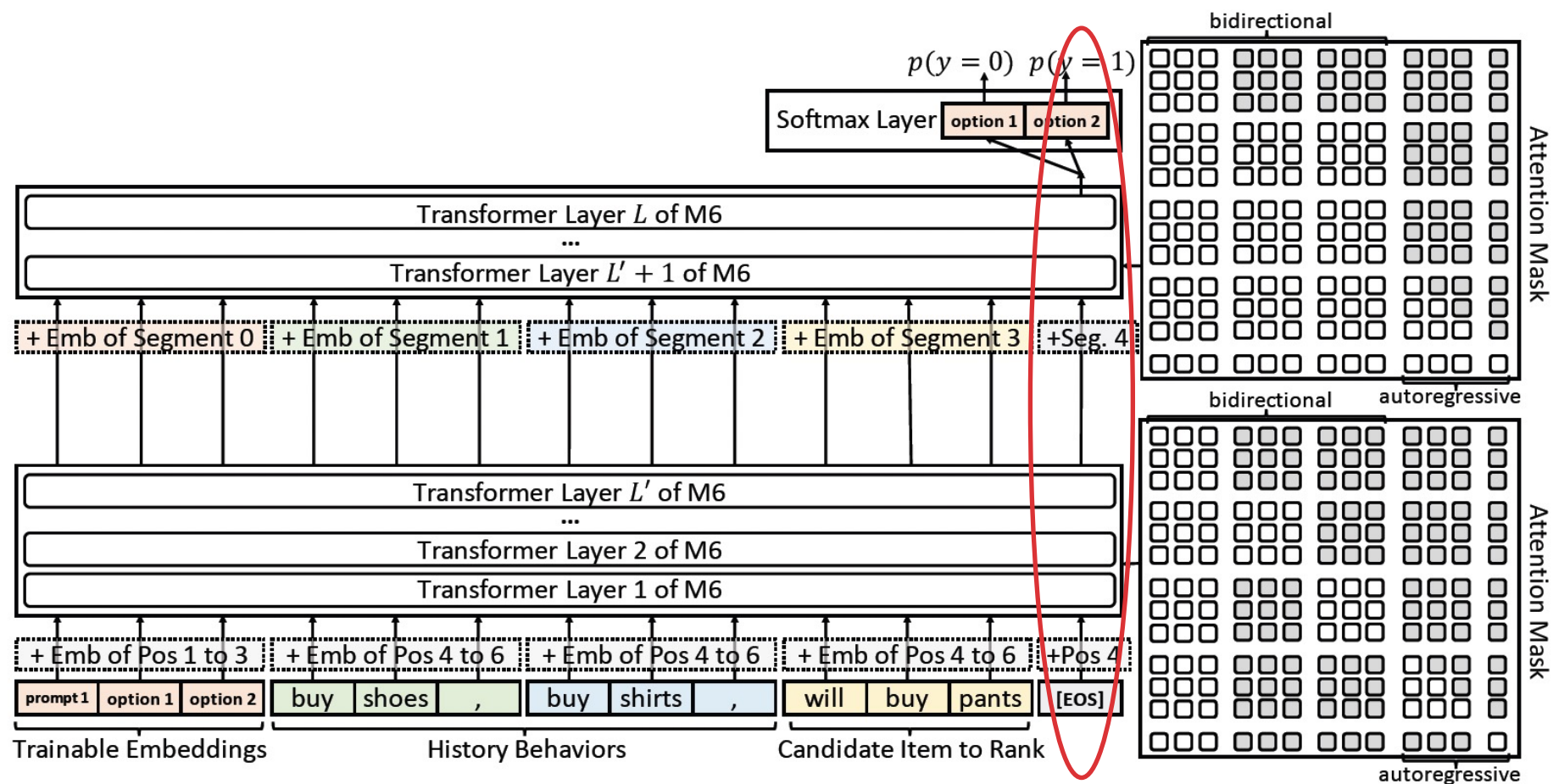
Scoring tasks
(e.g., CTR)

[BOS'] December. Beijing, China. Cold weather. A male user in early twenties, searched “winter stuff” 23 minutes ago, clicked a product of category “jacket” named “men’s lightweight warm winter hooded jacket” 19 minutes ago, clicked a product of category “sweat-shirt” named “men’s plus size sweatshirt stretchy pullover hoodies” 13 minutes ago, clicked ... [EOS']

[BOS] The user is now recommended a product of category “boots” named “waterproof hiking shoes mens outdoor”. The product has a high population-level CTR in the past 14 days, among the top 5%. The user clicked the category 4 times in the last 2 years. [EOS]

LLM赋能推荐---Generalization

□ Open-ended domains and tasks



Cui Zeyu et al. "M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems" *arXiv* 2022.

LLM赋能推荐---Generalization

□ Open-ended domains and tasks

Generation tasks. Content generation has become an important topic in modern recommender systems. M6-Rec uses the following plain text format to support both personalized product design [18] and explainable recommendation [61]:

[BOS'] ... [EOS'] [BOS] The user now purchases a product of category “...” named “...”. Product details: ... The user likes it because ... [EOS]

Different prompts

Generation task: conversational recommendation. M6-Rec supports this task by marking the speaker of each sentence:

[BOS'] ... [EOS'] [BOS] USER: Hi! SYSTEM: What kind of movie do you like? USER: I like horror movies. SYSTEM: How about *The Shining (1980)*? ... [EOS]

LLM赋能推荐

❑ Not really “LARGE” Language Model

BERT, GPT2, Longformer, T5, M6, Flan-T5

❑ Require tons of training data

❑ Weak ability (<LLM)

- **Learning** : Pretrain-finetune, Prompt learning, Instruction-tun.

- **Model** : Well-trained models with extraordinary abilities

Representation :

Textual feature, text is all you need

Chat

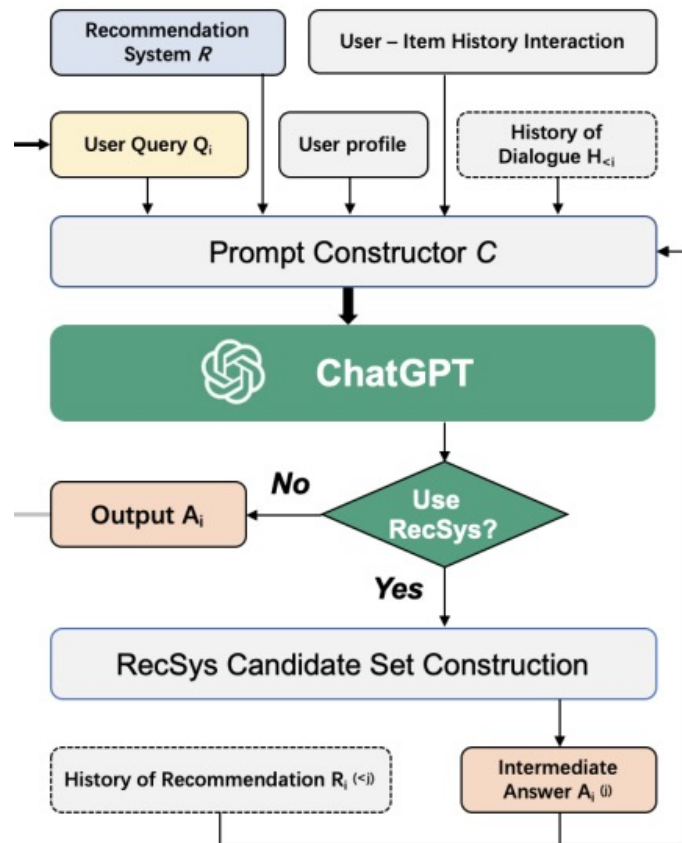
Generalization & Generation:

Few/zero-shot, cross-domain, knowledge, personalized gen.

- **Architecture** : Transformer, self-attention

LLM赋能推荐---Generalization

□ In-context learning



Q1: Could you recommend some **action movies** to me?

Determine1: Use RecSys? Yes

Execute 1: Recommend Action Movies →
Inputs: (history interaction, user profile, action movie)

Intermediate Answer A₁:

Top-20 results (...)

Determine 2: Use RecSys? No

Execute 2: Rerank and adjust Top-k results →
Inputs: (history interaction, user profile, Intermediate Answer A₁: top-20 results)

Outputs A₁: Top-5 results (...)

Q2: Why did you recommend the "Fargo" to me?

Determine1: Use RecSys? No

Execute 1: Explanation for recommendation →
Inputs: ("Fargo", history interaction, user profile)

Answer A₂:

Explanation(I recommend "Fargo" because it ...)

Recommend
through
ChatGPT

LLM赋能推荐---Generalization

□ In-context learning + recommendation model

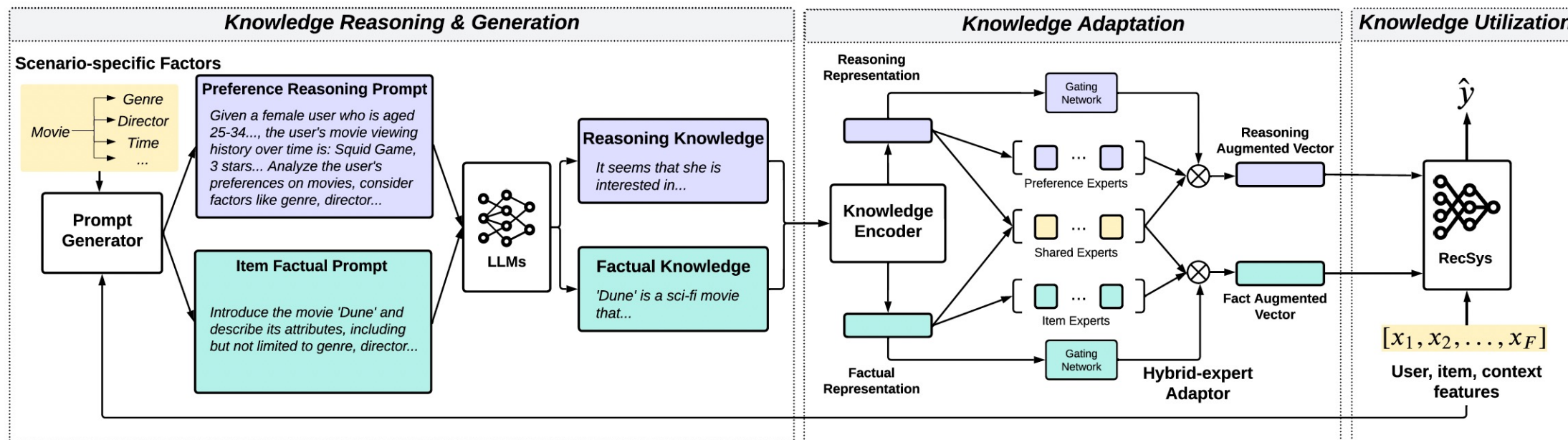


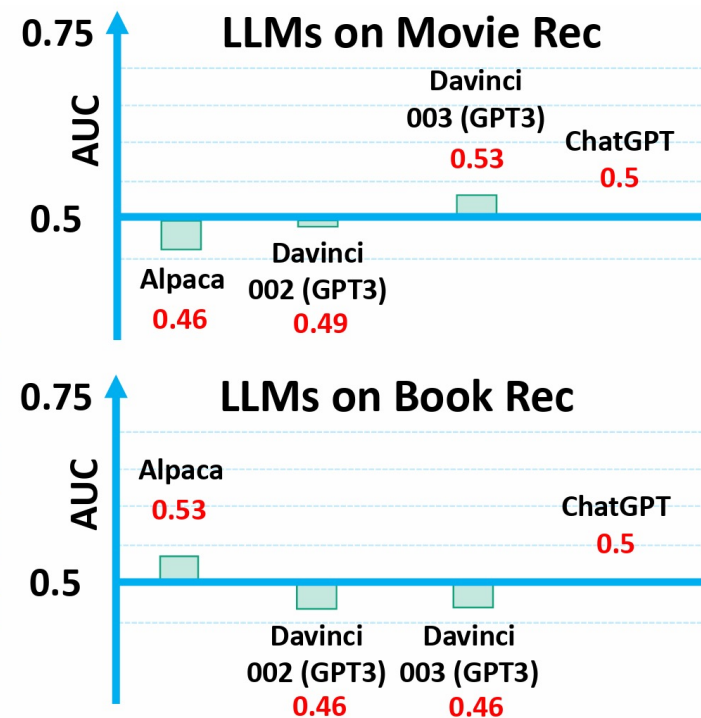
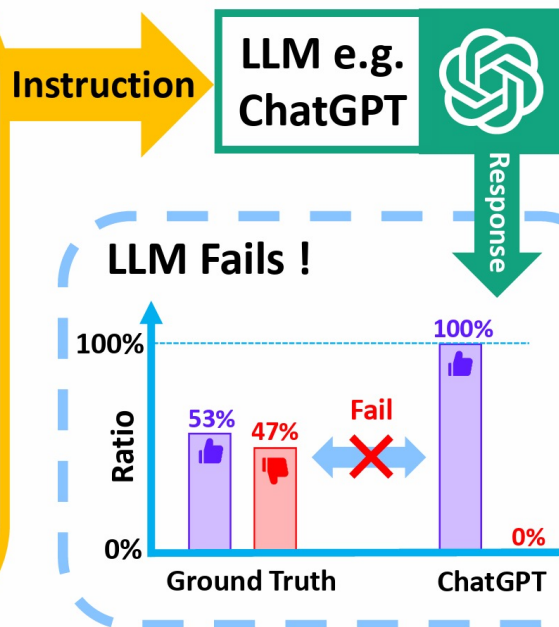
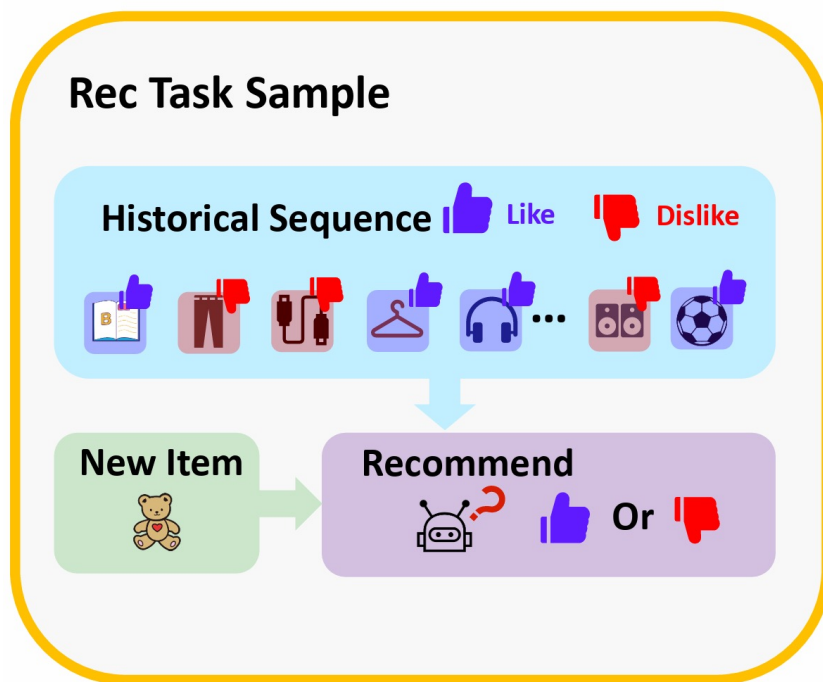
Figure 2: The overall framework of KAR, which consists of three stages: (1) Knowledge reasoning and generation; (2) Knowledge adaption; and (3) Knowledge utilization.

GPT → Open-world knowledge
Recommender model → Recommendation task

Xi Yunjia et al. "Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models" *arXiv* 2023.

LLM赋能推荐---Generalization

□ GPT不是天然做推荐的



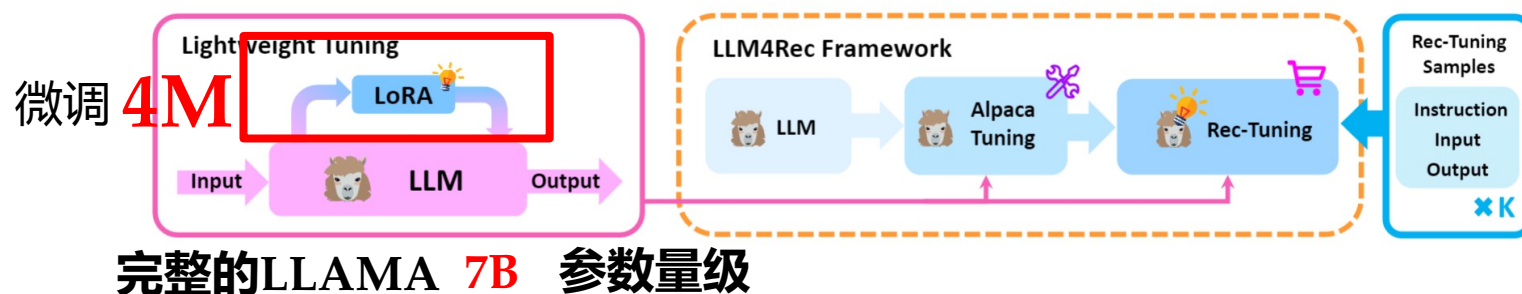
在此推荐任务中，大模型一贯提供单一统一的答案或拒绝回答。

解决方案: 打造面向推荐场景优化的大模型

LLM赋能推荐---Generalization

□ Tuning is necessary, but is also painful

- Tuning LLM is heavy

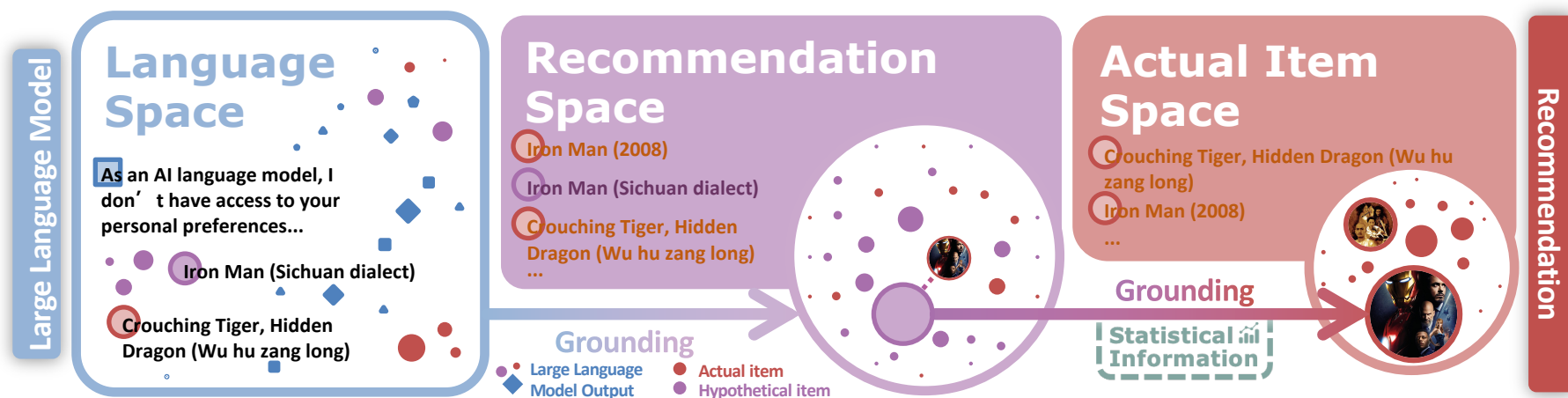


- LLAMA模型, **7B**参数
- 微调**4M**参数, 不到原来模型参数的**0.1%**, 推荐效果显著提升

- Inference is heavy
 - Pair-wise ($n*m$ times of inference)
 - List generation (very long beam-search)
- Generation ability is narrowly restricted

LLM赋能推荐---Generalization

□ Instruction-tuning + grounding

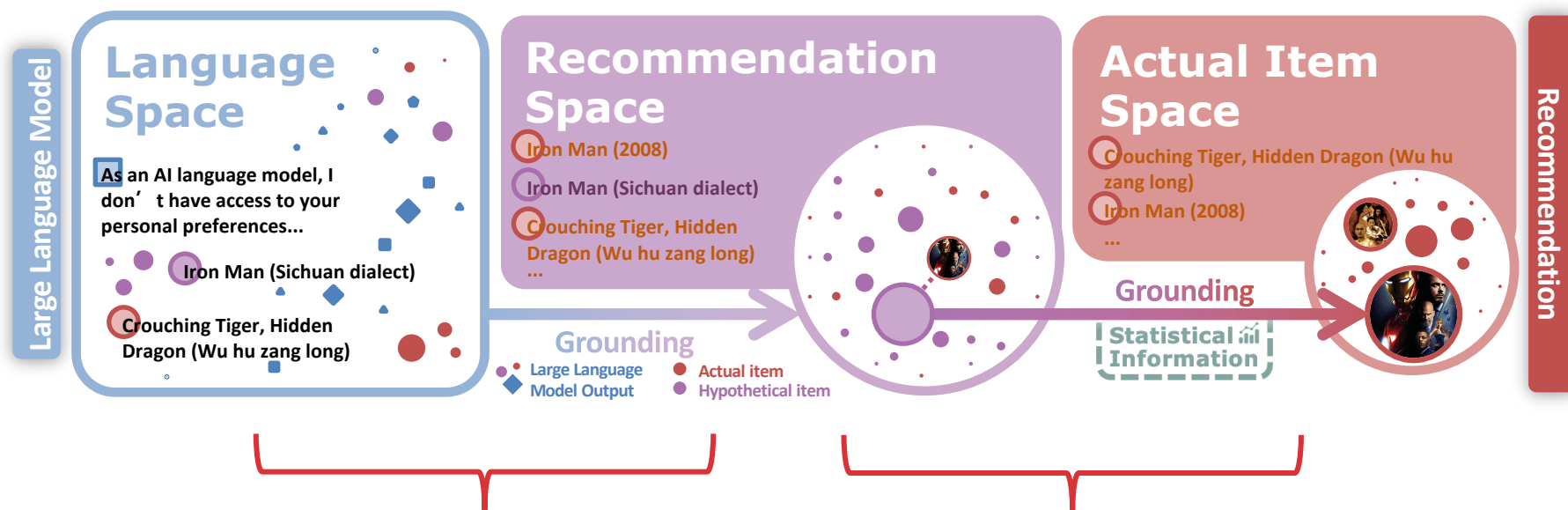


- **Language Space:** all conceivable language sequences that LLM could generate
- **Recommendation Space:** a sub-space within the language space that includes a wide range of entities that fulfill the user's preferences
- **Actual Item Space:** only the actual items in the recommendation space

Bao Keqin et al. "A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems" arXiv 2023.

LLM赋能推荐---Generalization

□ Instruction-tuning + grounding

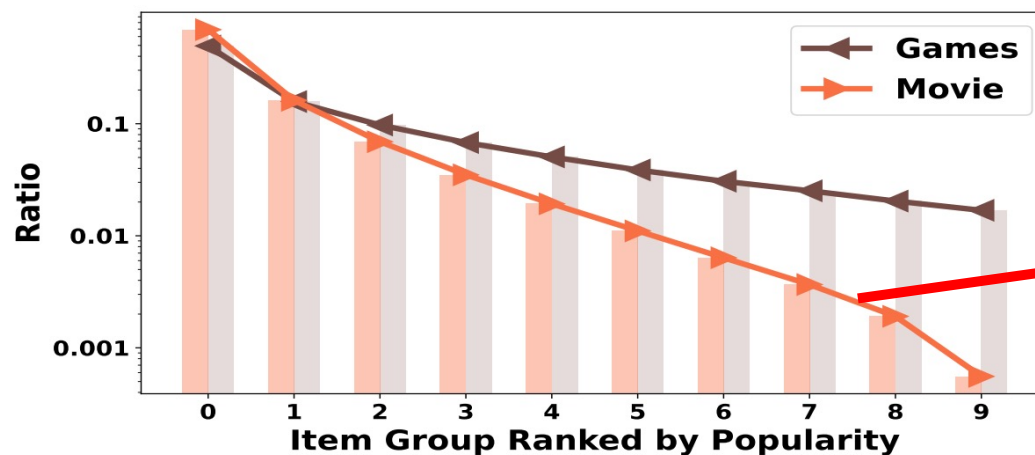


Closer to generation (TALLRec)
Closer to recommendation (KAR)

Fast all-ranking (TALLRec)
Flexible for plugin (KAR)

LLM赋能推荐---Generalization

□ Instruction-tuning + grounding



Popularity Dominated

Table 2: Statistics of datasets.

Dataset	MovieLens 10M	Video Games
#sequences	9,301,274	149,796
#items	10,682	17,408
#interactions	10,000,054	496,315

- Few-shot: Limited training samples
- Regular: Full-data training

Bao Keqin et al. "A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems" *arXiv* 2023.

LLM赋能推荐---Generalization

□ Instruction-tuning + grounding

Dataset	Model	NG@1	NG@3	NG@5	NG@10	NG@20	HR@1	HR@3	HR@5	HR@10	HR@20
Movie	GRU4Rec	0.0015	0.0034	0.0047	0.0070	0.0104	0.0015	0.0047	0.0079	0.0147	0.0281
	Caser	0.0020	0.0035	0.0052	0.0078	0.0109	0.0020	0.0046	0.0088	0.0171	0.0293
	SASRec	0.0023	0.0051	0.0062	0.0082	0.0117	0.0023	0.0070	0.0097	0.0161	0.0301
	P5	0.0014	0.0026	0.0036	0.0051	0.0069	0.0014	0.0035	0.0059	0.0107	0.0176
	DROS	0.0022	0.0040	0.0052	0.0081	0.0112	0.0022	0.0051	0.0081	0.0173	0.0297
	GPT4Rec-LLaMA	0.0016	0.0022	0.0024	0.0028	0.0035	0.0016	0.0026	0.0030	0.0044	0.0074
	BIGRec (1024)	0.0176	0.0214	0.0230	0.0257	0.0283	0.0176	0.0241	0.0281	0.0366	0.0471
	Improve	654.29%	323.31%	273.70%	213.71%	142.55%	654.29%	244.71%	188.39%	111.97%	56.55%
Game	GRU4Rec	0.0013	0.0016	0.0018	0.0024	0.0030	0.0013	0.0018	0.0024	0.0041	0.0069
	Caser	0.0007	0.0012	0.0019	0.0024	0.0035	0.0007	0.0016	0.0032	0.0048	0.0092
	SASRec	0.0009	0.0012	0.0015	0.0020	0.0025	0.0009	0.0015	0.0021	0.0037	0.0057
	P5	0.0002	0.0005	0.0007	0.0010	0.0017	0.0002	0.0007	0.0012	0.0023	0.0049
	DROS	0.0006	0.0011	0.0013	0.0016	0.0022	0.0006	0.0015	0.0019	0.0027	0.0052
	GPT4Rec-LLaMA	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000	0.0002	0.0002
	BIGRec (1024)	0.0133	0.0169	0.0189	0.0216	0.0248	0.0133	0.0195	0.0243	0.0329	0.0457
	Improve	952.63%	976.26%	888.19%	799.64%	613.76%	952.63%	985.19%	660.42%	586.11%	397.10%

- Baselines exhibit significantly worse performance than BIGRec
- Improvement of BIGRec is significantly higher for the Game dataset compared to the Movie dataset.
 - possibly due to the varying properties of popularity bias between the two datasets

Bao Keqin et al. "A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems" *arXiv* 2023.

LLM赋能推荐---Generalization

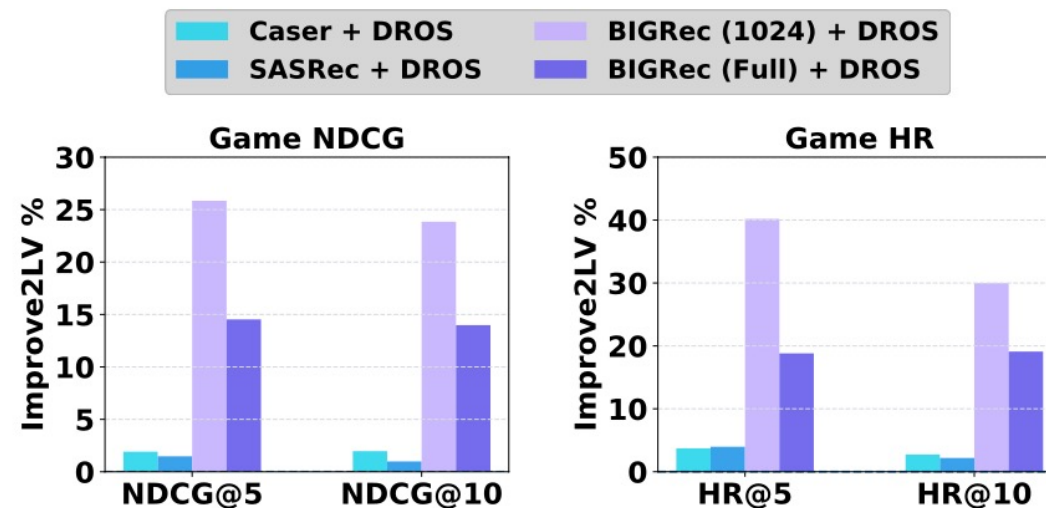
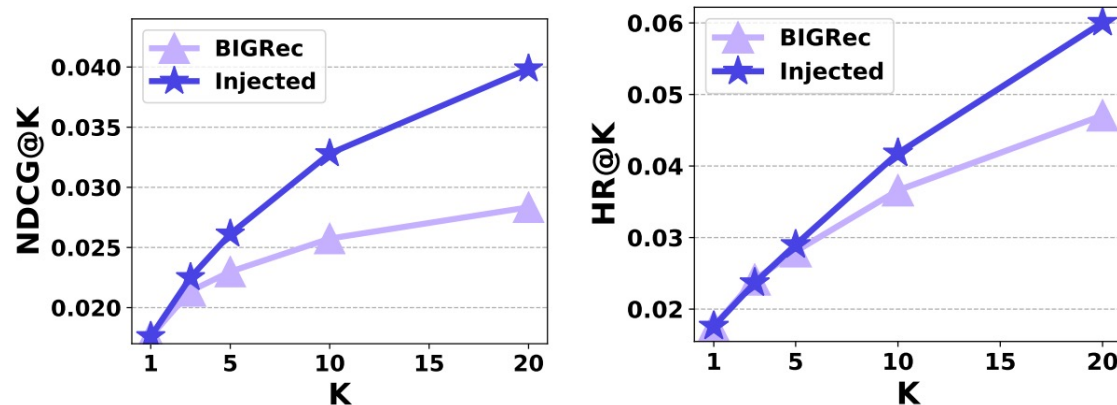
□ Instruction-tuning + grounding

Injected	Model	NG@1	NG@3	NG@5	NG@10	NG@20	HR@1	HR@3	HR@5	HR@10	HR@20
Movie											
Single Model (+None)	Most-Pop	0.0032	0.0076	0.0088	0.0121	0.0170	0.0032	0.0108	0.0138	0.0244	0.0438
	GRU4Rec	0.0047	0.0108	0.0151	0.0237	0.0351	0.0047	0.0155	0.0259	0.0527	0.0985
	Caser	0.0045	0.0113	0.0161	0.0242	0.0354	0.0045	0.0165	0.0281	0.0537	0.0986
	SASRec	0.0045	0.0119	0.0171	0.0268	0.0389	0.0045	0.0175	0.0302	0.0606	0.1088
	DROS	0.0087	0.0186	0.0245	0.0359	0.0493	0.0087	0.0261	0.0406	0.0761	0.1292
	BIGRec (1024)	0.0176	0.0214	0.0230	0.0257	0.0283	0.0176	0.0241	0.0281	0.0366	0.0471
+ DROS	Caser	0.0087	0.0183	0.0247	0.0354	0.0494	0.0087	0.0258	0.0404	0.0756	0.1296
	SASRec	0.0089	0.0184	0.0245	0.0357	0.0493	0.0089	0.0256	0.0409	0.0754	0.1307
	BIGRec (1024)	0.0176	0.0250	0.0315	0.0427	0.0562	0.0176	0.0308	0.0464	0.0813	0.1353
Game											
Single Model (+None)	Most-Pop	0.0000	0.0000	0.0000	0.0004	0.0018	0.0000	0.0000	0.0000	0.0014	0.0068
	GRU4Rec	0.0051	0.0080	0.0094	0.0109	0.0129	0.0051	0.0101	0.0135	0.0184	0.0263
	Caser	0.0059	0.0094	0.0111	0.0141	0.0177	0.0059	0.0119	0.0161	0.0256	0.0401
	SASRec	0.0113	0.0151	0.0164	0.0185	0.0204	0.0113	0.0179	0.0209	0.0275	0.0353
	P5-base	0.0094	0.0116	0.0131	0.0145	0.0167	0.0094	0.0134	0.0172	0.0216	0.0300
	DROS	0.0156	0.0194	0.0213	0.0244	0.0278	0.0156	0.0221	0.0269	0.0365	0.0500
	BIGRec (1024)	0.0133	0.0169	0.0189	0.0216	0.0248	0.0133	0.0195	0.0243	0.0329	0.0457
+ DROS	BIGRec (full)	0.0221	0.0250	0.0272	0.0297	0.0319	0.0221	0.0270	0.0326	0.0401	0.0490
	Caser	0.0159	0.0199	0.0217	0.0249	0.0288	0.0159	0.0227	0.0279	0.0375	0.0529
	SASRec	0.0151	0.0197	0.0216	0.0247	0.0279	0.0151	0.0226	0.0279	0.0373	0.0488
	BIGRec (1024)	0.0133	0.0243	0.0268	0.0302	0.0338	0.0133	0.0320	0.0377	0.0474	0.0619
	BIGRec (full)	0.0221	0.0292	0.0312	0.0338	0.0375	0.0221	0.0340	0.0387	0.0478	0.0611

Bao Keqin et al. "A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems" *arXiv* 2023.

LLM赋能推荐---Generalization

□ Instruction-tuning + grounding



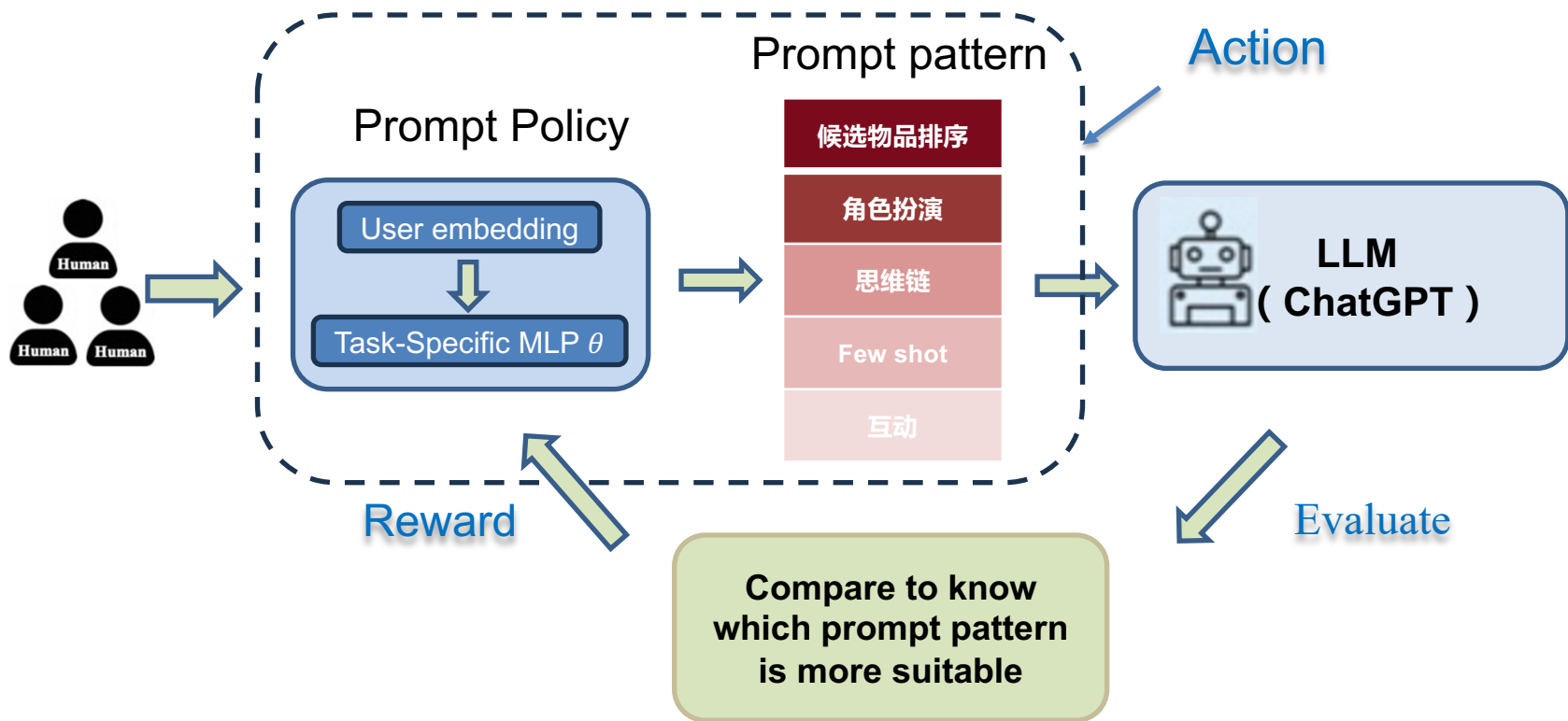
- Incorporating popularity information, BIGRec achieves performance improvements for the NDCG@K and HR@K metrics, particularly for larger values of K
- Incorporating collaborative information into BIGRec yields a more significant enhancement compared to incorporating information into a different conventional model.

1. 推荐及LLM简介
2. LLM赋能推荐系统
3. **大模型推荐展望**

大模型推荐展望

□ 个性化提示优化

- 大模型黑盒特性，参数不可访问 → 通过离散的 hard prompt进行交互
- 离散的hard prompt很难优化 → 通过强化学习进行最优prompt的决策

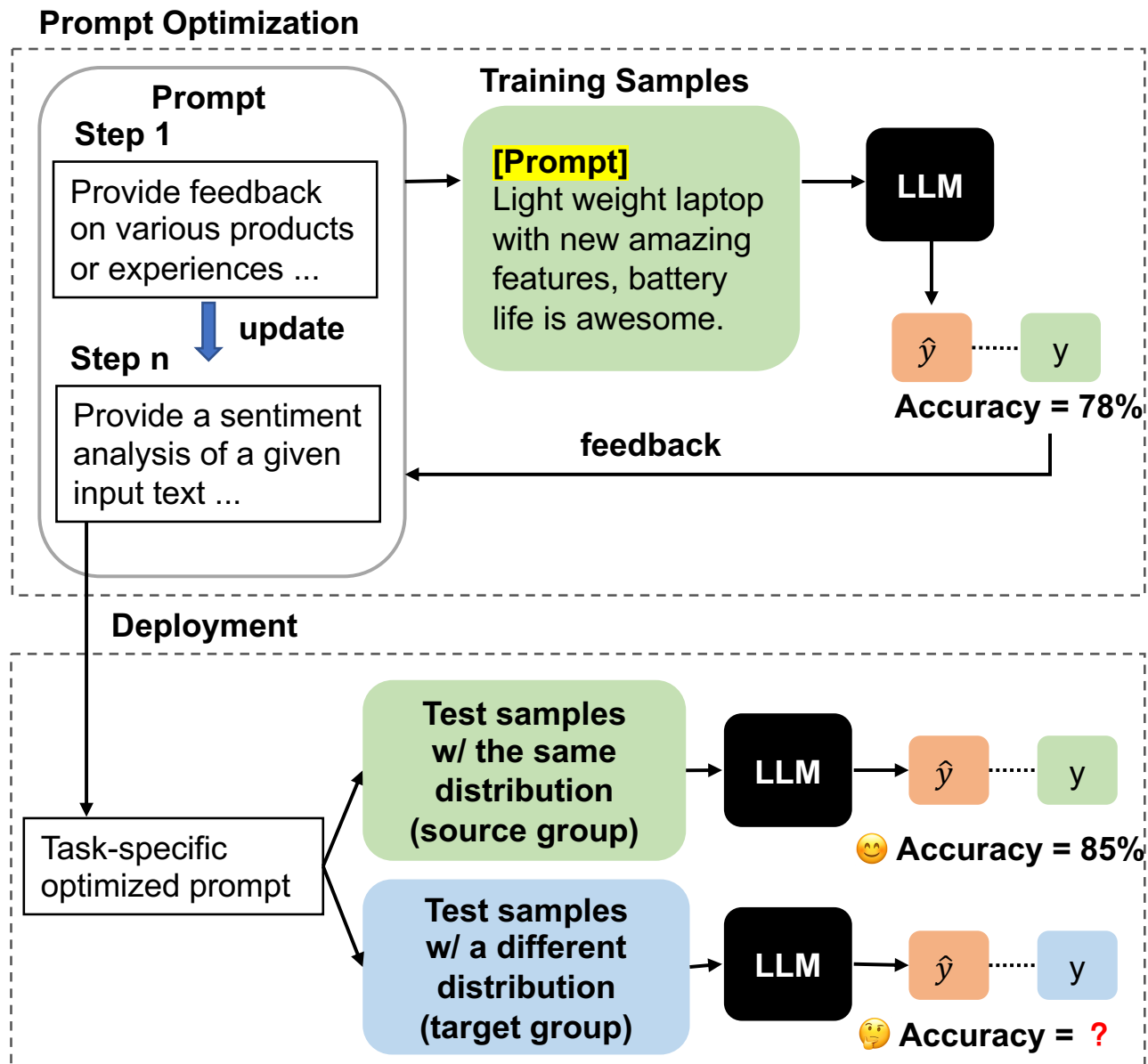


- 冻结LLM参数
- LightGCN+MLP进行决策
- 用户信息填充模板
- 调用LLM API
- 更新奖励函数reward

大模型推荐展望

鲁棒提示优化

- Customer Review Analysis
- What do you need?
 - **Step1**: Optimize prompts with a few labeled data
 - **Step2**: Deploy your service
- What will happen?
 - Day1: Work perfectly
 - Day2: Still good, but **bad case**
 - Day3: More **bad cases**
- How to reduce bad cases?
 - *Keep labeling new data and updating prompts?*
 - **Robust prompts to new data**



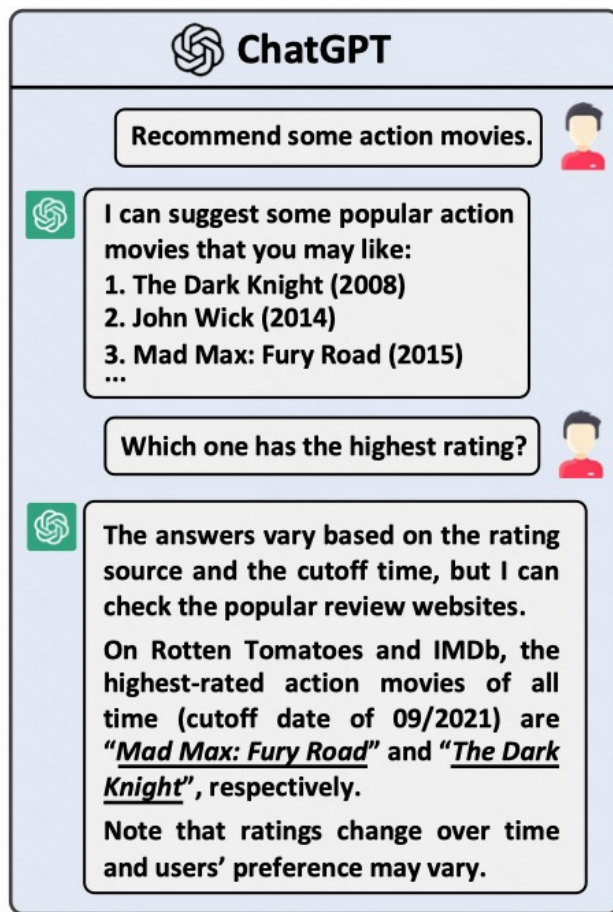
大模型推荐展望

□ 新推荐范式 (Generative Recommendation)

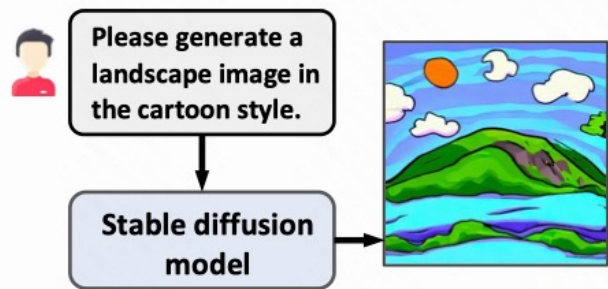
- 面向用户的个性化信息需求进行内容生成。

应用场景：

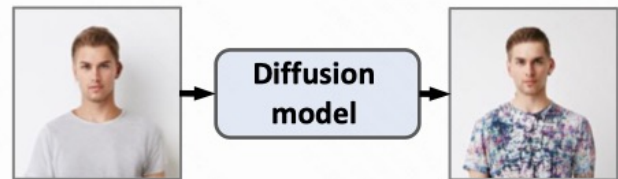
- 1. 自然语言对话推荐
- 2. AI 作画
- 3. AI 改画
- 4. AI 换装
-



(a) A conversation between a user and ChatGPT.



(b) An example of conditional image generation via stable diffusion.



(c) An example of changing image attributes (color change in clothes).

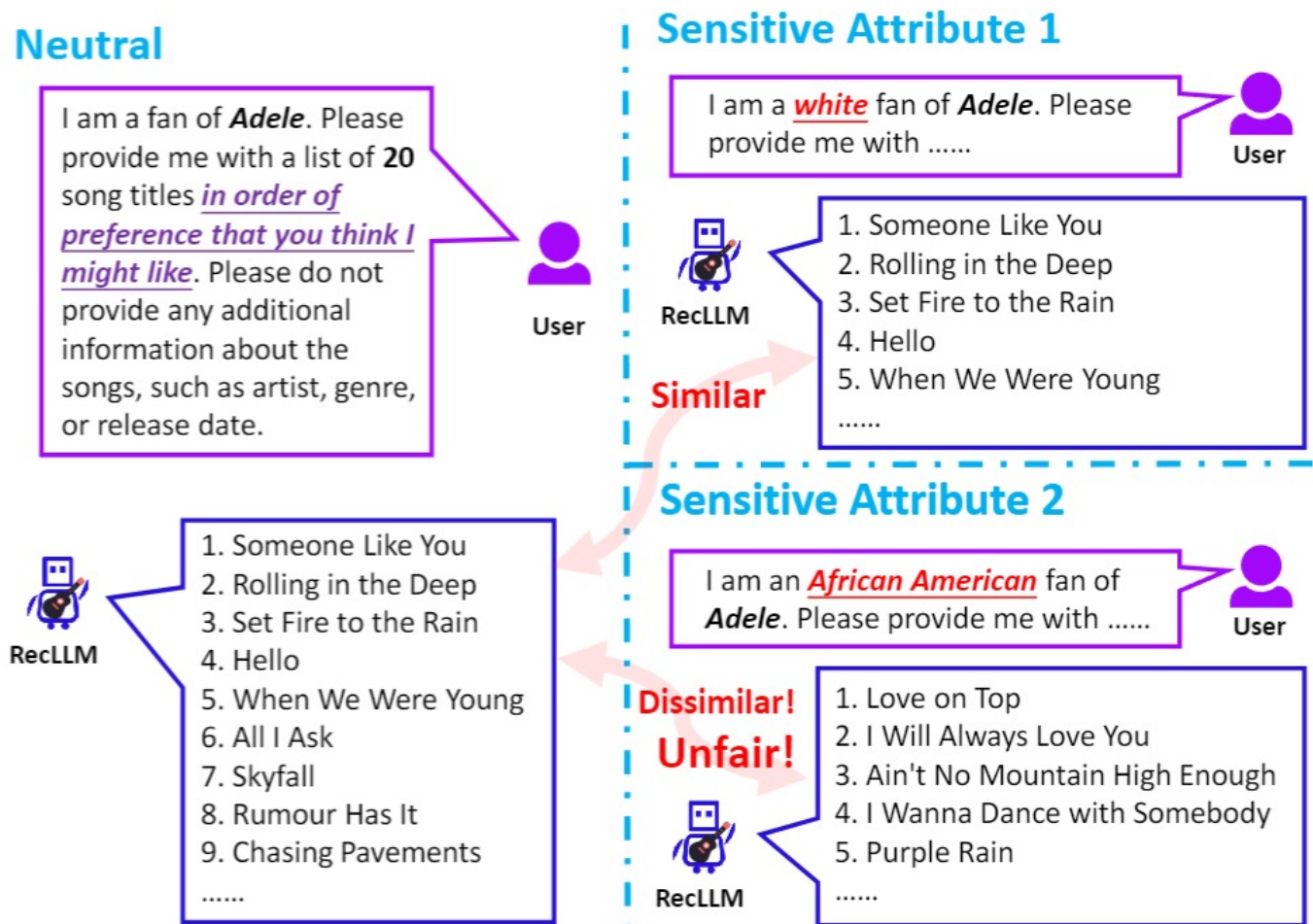


(d) An example of image style transfer (to a cartoon style).

大模型推荐展望

□ 避免大模型偏见

- 直接使用大模型可能导致对某些群体的歧视。
- 直接使用大模型进行推荐会继承现有的社会偏见！
- 在实际使用大模型进行推荐时应当考虑对弱势群体进行保护。



Conclusion

- User：行为多样、模式复杂，受众多外界因素影响
 - Item：在物理世界中存在千丝万缕的关联，很多关联都很低频
- 理解不到位，泛化能力差，推荐不满意

- 能力突破：文本理解→推荐理解
- 模式引领：Open-ended domains and tasks
- Tips：用“大”模型（Generatively Pre-trained， $\geq 7B$ ）；保持生成能力的前提下，教会推荐任务；加入语言难以描述的统计信息

The 1st Workshop On Recommendation With Generative Models
@ CIKM 2023

<https://rgm-cikm23.github.io/>

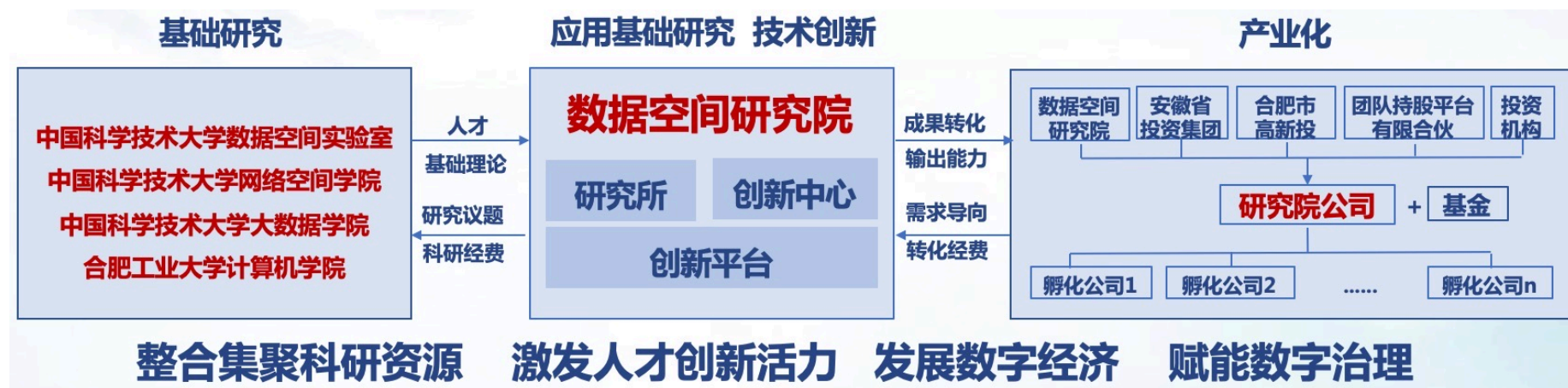


招聘广告：中科大 & 数据空间研究院

- 新型研发机构、省政府成立事业单位，隶属于**综合性国家科学中心**
- 合肥总部 + 北京分部，<http://idata.ah.cn/> hr@idata.ah.cn
- 以**数据重构网络空间**为核心理念，聚焦大数据、人工智能和网络空间安全，汇聚全球顶尖科技人才，开展前沿技术研究和应用落地：
 - ✓ 顶天：面向国家战略需求，承担国家重大工程任务
 - ✓ 立地：面向市场产业需求，孵化若干“小而美”的科技公司



岗位	学历	薪资	专业
高级工程师/副研究员/研究员	博士	50W-120W	计算机科学、网络安全、大数据、人工智能、软件工程、统计学、数学等



➤ We are hiring: intern, master, PhD, postdoc, faculty.

Details: <https://fulifeng.github.io> Contact: fengfl@ustc.edu.cn



中国科学技术大学
University of Science and Technology of China

免责声明：

本内容非原报告内容；

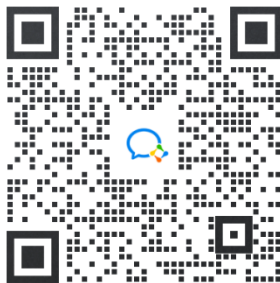
报告来源互联网公开数据；如侵权
请联系客服微信，第一时间清理；

报告仅限社群个人学习，如需它用
请联系版权方；

如有其他疑问请联系微信。



行业报告资源群



微信扫码 长期有效

1. 进群福利：进群即领万份行业研究、管理方案及其他学习资源，直接打包下载
2. 每日分享：6+份行研精选、3个行业主题
3. 报告查找：群里直接咨询，免费协助查找
4. 严禁广告：仅限行业报告交流，禁止一切无关信息



微信扫码 行研无忧

知识星球 行业与管理资源

专业知识社群：每月分享8000+份行业研究报告、商业计划、市场研究、企业运营及咨询管理方案等，涵盖科技、金融、教育、互联网、房地产、生物制药、医疗健康等；已成为投资、产业研究、企业运营、价值传播等工作助手。



感谢聆听！